Finding Coincident Data From Satellites: Using “Meta-Metadata” To Reduce Load on Archive[1]

Ted Willard
Computer Sciences Corporation
Lanham-Seabrook, Maryland

John Berbert
Goddard Space Flight Center
Greenbelt, Maryland

Abstract

In this paper, the concept of “meta-metadata” is introduced as an abstraction of metadata as metadata is an abstraction of the data. The problem of finding coincident data from satellites is examined using metadata and meta-metadata. The meta-metadata approach is shown to be more generally applicable and supports a more heterogeneous system. Finally, the implementation of a prototype system for performing coincidence search is described.

The challenges and opportunities for archives are evolving. The number and scale of archives are increasing, as is the interoperability of archives through the use of standards and Internet connectivity. Archivists are challenged because the demands on archives are increasing faster than the technology; however, there is great opportunity because of the large and diverse volume of data online. One way to reduce the challenge and to take advantage of the opportunity is through the introduction of meta-metadata. Metadata reduces the demands on an archive by providing information about archive holdings in a more compact form than is available through the data, and it may support the discovery of related data in an archive. Meta-metadata reduces the demands on archives by representing the metadata in a more compact form, and it may support the discovery of related data across multiple archives. The ability to search across multiple archives becomes valuable as the variety of data online allows combining data in ways not envisioned when the data were collected.

The problem of finding coincident data from satellites can be exemplified by two questions: (1) When does a specific instrument onboard a specific spacecraft view a particular wheat field in Kansas?, and (2) When do two specific instruments on two spacecraft both view a particular wheat field in Kansas within half an hour of each other? The National Aeronautics and Space Administration’s (NASA’s) Earth Observing System (EOS) Data and Information System (EOSDIS) collects spacecraft data from multiple spacecraft and supports the interdisciplinary use of those data. In some cases, however, the metadata provided with the data does not facilitate coincidence searches. EOSDIS metadata for spacecraft data ignore the inherent relationship between the space and time dimensions of the data. For data granules with large temporal extent, ignoring the space-time relationship leads to spatial metadata with little or no information content. The lack of information in the metadata leads to coincidence queries that are all hits or all misses. To perform space-time coincidence queries on such metadata can be a resource-prohibitive endeavor because of the large numbers of false-positive data granule hits. Once the relevant data granules are found in the database by eliminating the false-positive hits, it is frequently beneficial to subset those data granules to obtain the data of interest only. To spatially subset data that is organized by time of data collection, space must be related to time.

The natural way to relate space and time for spacecraft data is through an orbital model that enables the space-time relationship to be accurately represented by eliminating false-positive database hits and the times of coincidence to be used to subset the data. Converting spatial queries into temporal queries could substantially reduce the load on large geospatial databases.

The Coincidence Search Prototype (CSP) was designed to use models to provide coincidence search services for the EOSDIS Core System (ECS) and the related Version 0 (V0) data servers. The CSP does not rely on the spatial metadata, but uses orbital data available on the Web and spacecraft attitude and instrument models to convert coincidence queries in space and time to queries only in time. Since time-based queries generated using the orbital model do not generate the false-positive hits that spatial queries generate on the metadata, the data volume is reduced to the data granules that contain relevant data only. The times of coincidence may be used to subset the individual data granules to further reduce the data volume transferred from the archive to the end user.

[1] This work funded by the ESDIS Project (GSFC Code 423).

Introduction

The metadata is not always enough—and it is the archive and the end user that pay the price. If the metadata does not provide sufficient information for a user to identify the relevant data only, the user is forced to retrieve the irrelevant with the relevant and to weed out the irrelevant data. This retrieval and weeding process places an undesirable load on the user and the archive that should be eliminated by identifying the meta-metadata that provides the information the user needs in a compact form.

In the case of the ECS, the users required the system to identify coincident data. The ECS implementers were unable to provide a coincident search capability in part because, as it will be shown in this paper, the metadata were insufficient. It will be further shown that forcing users to request the irrelevant with the relevant data can increase the load on the archive by two orders of magnitude. Trying to force the data providers to provide the necessary metadata may be a lost cause and may not be desirable.

The CSP was the solution for ECS—because it looked beyond the metadata to information that was more abstract than the metadata. Orbital, attitude, and sensor models were used to provide an expectation of the data to be found in the metadata. Since the models provide data about the metadata, we called them meta-metadata. The use of meta-metadata provides not only faster and more comprehensive search capabilities at a single archive, but potentially supports cross-site searches. The CSP is separate from the metadata database and communicates using the standard protocol, Object Design Language (ODL), so that the CSP is able to search not just one archive but multiple archives—hence, the CSP is an archive search engine.

This paper is organized into this introduction, a background section, a description of the limitations of a metadata-based approach (catalog search) to a spatial-temporal coincidence search, a description of how meta-metadata works for spatial-temporal coincidence search, a description of the coincidence search prototype, a summary, a glossary, an explanation of calculations in the paper, and references.

Background

This section provides the context for the CSP. It defines the problem of coincidence search for satellite data and two classes of metadata, wholesale and retail, and it describes why data producers tend to favor wholesale metadata, while data users prefer retail metadata. It is explained why wholesale metadata is inadequate for spatial-temporal coincidence search, why this is of concern to the archivist, and why the archivist needs to provide data at a retail level to the users. The potential for multiple retailers of data is described, as well as the potential for a single retailer to support multiple archives to support cross-archive coincidence searches.

Satellite coincidence search

The purpose of the CSP is to find spacecraft data that is coincident in space and time with something else, including other data. For example, “What data are available when 2 instruments on 2 different spacecraft view Kansas within 6 hours of each other?”, or “What data are available from one instrument when viewing Kansas?” The capability is important to NASA because ECS is supposed to store not only earth science data, but support interdisciplinary research. Interdisciplinary research requires combining data from different instruments on different spacecraft to produce new knowledge. Since the data come from spacecraft in orbit, there is a fundamental relationship between the time of data collection and the spatial location of the quantities being measured by the instruments. This fundamental relationship provides the basis for the meta-metadata.

Wholesale and retail metadata

The metadata provided for earth science data can be described as either wholesale (will not support coincidence search) or retail (will support coincidence search). (The exact separation between retail and wholesale is fuzzy, but a retail product spans only a fraction of an orbit. For ECS, products that span 15 minutes or less would be considered retail, 2 hours or more would be wholesale, and in-between would be debatable.) Frequently, earth science data providers produce their data in a form for efficient generation of higher-level data products rather than in a form convenient for other users. Dr. Bruce Barkstrom [1] refers to this data production as the “wholesale” approach, and we have adopted his terminology. Barkstrom describes users as wanting a retail approach to data, data producers as taking a wholesale approach, and archivists as the mediators between the two. Meta-metadata provides the archivist with a way to mediate between the retail demands of the consumer and the wholesale demands of the producers.

Wholesale metadata inadequate for coincidence search

Wholesale metadata is inadequate for coincidence searches because the spatial metadata for wholesale data generally contains little or no information. Generally, the spatial extent of the data product is recorded in metadata as the latitude and longitude extremes of the data. In the case of a product that spans an orbit or more, the spatial domain of the data set spans the range of latitudes and longitudes over which the instrument views. In such a case, every product has the same spatial metadata: the latitude and longitude extremes of the instrument. Since all the products have the same spatial metadata, the metadata has no information content. Additionally, the metadata is deceptive; for example, an instrument with a small field of view will only view a small fraction of the Earth in one orbit, although the metadata might report that the whole Earth had been viewed. From the metadata alone, it is not possible to determine which data products include views of an area target such as Kansas. Since all the data products have the same metadata, all or none of the data products will be reported as viewing Kansas. Presumably, some fraction of them will view Kansas, and somehow it must be determined which products view Kansas. The worst option is to require the user to order all the data and then sort the wheat from the chaff.

Wholesale demands of producers

Data producers require a pipeline to process the data received from an instrument as fast as it is received. Efficiency, consistency, and thoroughness are key. The data producer requires an efficient process to produce products that are consistent over the life of the mission. The requirement for efficiency stems from the need to process all the data without a growing backlog. For efficiency, long timespans (1 orbit or 1 day) are typically used for each product, sometimes at the cost of requiring dedicated systems to handle the large file sizes.

Retail demands of users

Users require flexibility rather than volume. In general, users do not want nor can they use the large files generated by the data producers. The users lack special facilities for handling the data products, so the very large files are difficult to use. In addition, the users do not need the large files because most of the data in a large file is irrelevant to a given user. For low-level products, the large files group data by time (for example, June 1), and users want data grouped in space (for example, Kansas). The user may need only a small subset of the data contained in each of many large files. For example, if the user is interested in Kansas, and the instrument views Kansas only once a day at most for about 3 minutes, large files sizes do not benefit the user. Regardless of whether the data were packaged in 1-day or 10-minute chunks, the user would order the same number of files. In the former case, however, two orders of magnitude more data were shipped to the user.

Archivist as retailer

The archivist must either live with demands on the archive that vastly exceed the users' needs or must develop tools to select the relevant data and do the subsetting. With the explosive growth in archive size, the first option is becoming increasingly prohibitive. The user might reject the first option because, having received an excessive volume of data, an army of graduate students is needed to find the relevant data. Both the archivist and the user gain by having tools at the archive site that enable the user to find relevant data rather than shipping the user a lot of chaff with the wheat. Meta-metadata is just such a tool.

Third-party retailers: archive search engines

The archive does not have to be its own retailer or be limited to a single retailer. The user needs to be able to find just the relevant data and, preferably, to reduce the size of the data sets transferred to the user to contain only the required data. For the user to be able to find the relevant data, the archive must either be able to act as a retailer or provide an interface to the archive for a third party to act as a retailer. If the archive provides an interface for third parties to act as retailers, different user communities can be served by different third parties. If multiple archives use the same interface for allowing third parties to act as retailers, a single third party can act as a retailer across multiple archives. A third party that searches multiple archives can be called an archive search engine. Conceptually, this is similar to search engines on the Web, such as Yahoo or Lycos. Users can select such a search engine to find applicable information from different Web sites rather than having to investigate each site individually.

Object Description Language as common archive interface

ODL is the common archive interface used by CSP; it was developed by the Jet Propulsion Laboratory (JPL) for its Planetary Data System (PDS). Both the currently operational EOSDIS V0 and the upcoming ECS respond to search queries formatted in ODL. The CSP is able to remotely query the V0 system and should be able to query the ECS when it becomes operational. Users will be able to find data that are coincident between ECS and V0, although the data will be stored in different archives. The CSP is therefore an archive search engine.

Limits of metadata-based spatial coincidence search

The alternative for performing coincidence search via an orbital model is to perform a catalog search. Using the metadata in the catalog, a user can attempt to find coincidences. If the data provider uses a wholesale approach to providing data, this approach will not be productive. If the data provider uses a retail approach, the method will be productive for coincidences with other retail data—but our belief is that it is not as efficient a method as the orbital model approach.

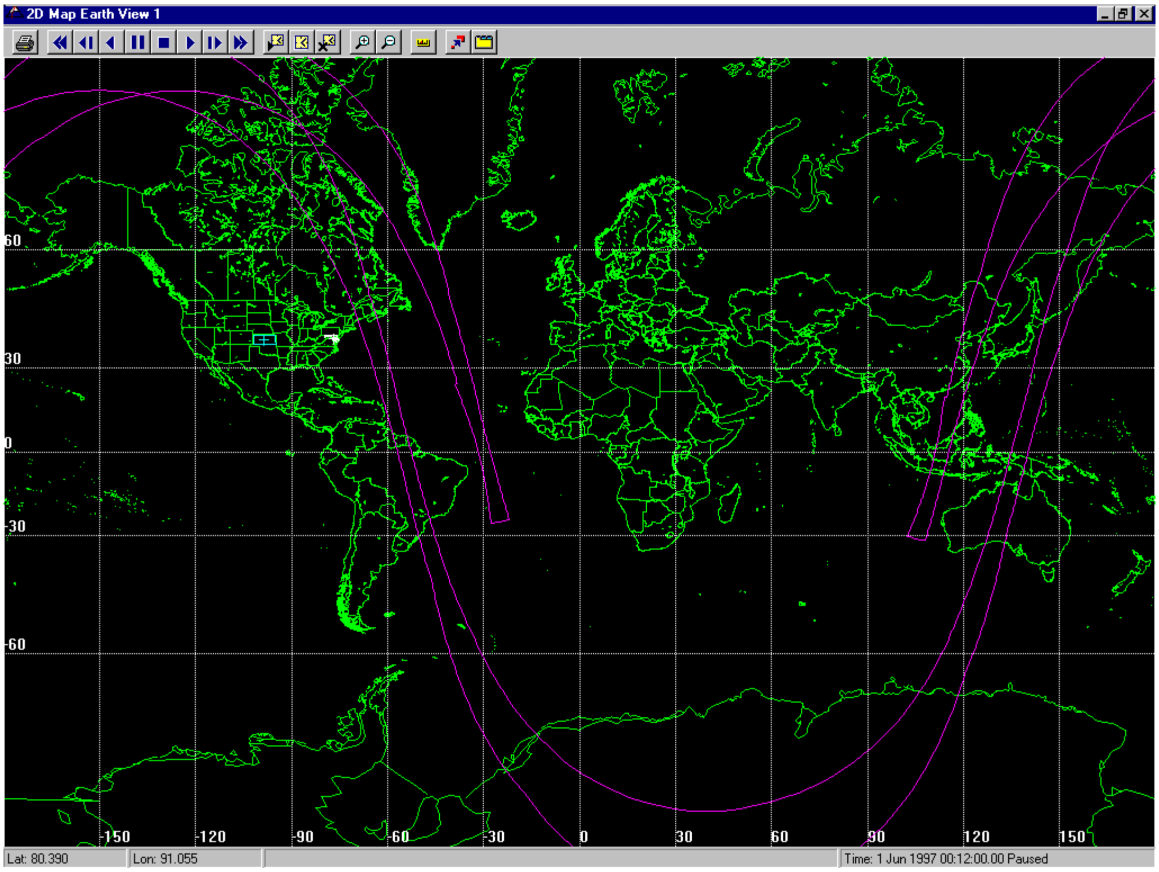
Problem with catalog search (wholesale)

If the data producer provides the data at the wholesale level, coincidence search via catalog search may not be possible. The problems can be understood better by looking at the example of the Multiangle Imaging Spectroradiometer (MISR).

MISR (pronounced “miser”) is an instrument onboard the EOS-AM1 spacecraft that is due to launch in 1999. MISR is a wholesale data source because it will generate products 2.5 hours in length and longer, which is more than 1 orbit period for EOS-AM1. Figure 1 shows a 2.5-hour swath corresponding to the MISR product. Note that the swath is wide enough to include the North and South Poles, and it includes data for the entire range of latitudes and entire range of longitudes.

To represent the extent of a 2.5-hour MISR product using a bounding box, one must specify the entire Earth. As is apparent in Figure 1, a 2.5-hour MISR product covers only a fraction of the Earth. The width of the swath is 360 km, and the swath is 60,800 km long [2.5 hrs/orbit period (=98.88 minutes) * circumference of the Earth]. For a spherical Earth, the area covered by the swath is 21,885,000 km^2, and the surface area of the Earth is 255,593,000 km^2, so the swath covers less than 8.6 percent of the Earth. (Actually, the swath covers less than 8.6 percent because it covers some parts of the Earth twice.) For a random point on the Earth, there is a less than 8.6 percent chance that the point is covered by any given MISR product; however, the metadata for every MISR product state that the product covers the entire Earth.

For a point on the Equator, the situation is even worse. The angle between the orbit plane EOS-AM1 and the Equator is 18.3°, so the projection of the swath onto the Equator is 342 km (=360 km * cos (18.3°)), or 3.07°. To cover the entire Equator would require at least 117.26 Equator crossings (=360°/3.07°), or 58.82 orbits (=equator crossings/2), or 97 hours (orbits * orbit period). (There are overlaps, so it actually takes even longer to cover the entire Equator.) Since MISR requires more than 97 hours to view the entire Equator, the probability that a given point on the Equator is included in a 2.5 hour product is less than 2.6 percent (=2.5/97).



Lat: 80.390

Lon: 91.055

Time: 1 Jun 1997 00:12:00.00 Paused

Figure 1: Multiangle Imaging Spectroradiometer (MISR) groundtrack.

Using catalog search to find coincidences in wholesale metadata does not work because the metadata does not distinguish one data product from another. The catalog search could be supplemented with code that reads the actual archive data, but this is inefficient and may not work.

A catalog search (metadata query) for MISR data covering any point on the Earth will return all MISR data, although less than 8.6 percent of the data will cover that point. A catalog search for MISR data at the Equator will also return all MISR data, although less than 2.6 percent of the MISR data will cover the point. If the user requests data in which a random point on the Earth is simultaneously viewed by both MISR and some other wholesale data provider with characteristics similar to MISR, a catalog search will return all the data for MISR and for the other instrument. The probability that the data are actually coincident is less than 0.74 percent (= (8.6 %)^2) or worse than 1 in 135. Without aid in finding the relevant data, the user must retrieve 135 MISR products and 135 products of the other instrument to find 1 pair of products that are coincident.

Supplementing the catalog search by retrieving and reading the archived data may work, but is inefficient. It would be inefficient to read in all the data for MISR whenever one performs a coincidence search because it would be as if no metadata existed at all. In addition, it would be difficult to implement and maintain because it requires detailed knowledge of the format of each MISR product and the other products in the archive and requires separate implementations for each product. It may not even work because the archived data may not include spatial information. Instrument data are telemetered to the ground with time-tags, but not necessarily spatial information. Spatial information may not be connected to the instrument data until after orbit data are generated from spacecraft tracking data and the instrument data are further processed.

Another problem with catalog search is that it does not support subsetting. MISR has multiple cameras for viewing along the ground track of AM-1. From the time when a point is visible beginning with the most forward-looking camera to the time when it is almost out of sight of the most rear-facing camera is 14 minutes. If the user is interested only in the time when MISR views the point on the Earth, then it would be preferable to subset the 2.5-hour product to a 14-minute product. This would reduce the data volume by 16 percent of its unsubsetted volume. (Much greater reductions would be possible if the user wanted only the data from the central camera because it has the least distortion.) The catalog does not relate the space and time attributes of the data and therefore provides no information to subset by time. Depending on how the data are organized, it may or may not be possible to spatially subset the data. Lower-level data products are organized by time of data collection, but spatial data may not be available when the products are generated. It is only in postprocessing that spatial data is added.

In conclusion, for MISR a reduction in data volume of two orders of magnitude can be gained by selecting only the data products for a pair of instruments that both view a spot on the Earth. An additional reduction by an order of magnitude is possible by subsetting the data before transmission to the user.

Problem with catalog search (retail)

While catalog search is not a viable option for coincidence search for wholesale data, it can be used for retail data. First, the data provider is required to provide the data in retail format, something over which the archivist has no control. It is also required that if there are coincidences between two sources of data that both data sources come from retail providers. Second, the metadata database must support geospatial queries—it must be possible to determine if the bounding area specified in the metadata intersects with the user’s target area. Third, if the archive is distributed like ECS, metadata from one metadata database must be transferred to the other database. Finally, it requires a large metadata database to hold the large volume of metadata.

The first requirement, receiving data in retail format, is met by ECS for instruments Moderate Resolution Imaging Spectroradiometer (MODIS) onboard AM-1 and ETM+ onboard Landsat-7. Many more products within ECS are not retail, so coincidences between them and the MODIS or Landsat products cannot be found with catalog search. MODIS data granules span 5 minutes (about 5 percent of an orbit). Landsat-7 data granules are catalogued as if the granules were about 30 seconds in duration (about 0.5 percent of an orbit) although they may, in fact, be 17 minutes long. The accuracy of the search is limited by the duration of the longer of the two products. Since MODIS products are 5 minutes long, any coincidence computation using catalog search has an inherent uncertainty of 5 minutes.

The second requirement—that the metadata database supports geospatial queries—increases the demands on the database. The capability to perform geospatial queries is available as an add-on capability of Informix Universal Server (IUS) and Sybase, which were the databases considered for ECS. Geospatial queries are more CPU intensive than temporal queries and therefore require more resources.

The third requirement for transferring metadata from one database to another only applies to distributed systems such as ECS. ECS does not have a single metadata database for all data stored in the archive, but has at least one database at each data center. To find coincidences between a MODIS level-1 product and a Landsat scene, one would have to search the metadata databases at the Goddard Space Flight Center (GSFC) for the MODIS product and at the Earth Resources Observation System (EROS) Data Center for the Landsat scenes and somehow combine the results.

Depending on the duration of the time covered by the search, and the size of the area of interest, the problem can be manageable or unwieldy (see Table 1). The first row of Table 1 shows that if the time duration is 30 days, and the size of the area of interest is the same as the area of Kansas; depending on the latitude of the area of interest, about 5.4 of the 86,400 scenes for the month would include the area (see “A Note on Calculations” below for an explanation of the calculations). For MODIS, approximately 82 of the 8,640 MODIS products would include Kansas. The metadata for the 5 or 6 Landsat products could be reasonably transferred to the MODIS metadata database, and 5 or 6 searches could be made against the 82 MODIS products to find if there were any that both viewed the same region of the area of interest at the same time.

Conversely, the third row of Table 1 shows that if the user wanted to find coincidences between two instruments, but did not care where they occurred in the course of a 30-day period, the user would have to load the metadata for 8,640 MODIS products into the Landsat metadata database and perform 8,640 (resource-intensive geospatial) searches against the 86,400 Landsat scene products. This would be a greater load than could be reasonably supported at most facilities.

In the midrange (second row of Table 1), if the user was interested in a region the size of the continental United States, 1335 searches could be made against the 8,640 MODIS products or 212 searches could be made against the 86,400 Landsat products. This problem grows by 2 orders of magnitude if the time length increases from 1 month to 10 months because the number of searches and the size of the data to be searched each grow by an order of magnitude. Therefore, while a search for a month may not be prohibitive for a site, many searches spanning longer timespans could easily overwhelm a site. [In practice, however, none of the searches—the Kansas sized, the whole Earth sized, or the continental U.S. search—would find any data because instruments are onboard spacecraft whose orbits are such that one spacecraft follows the other one. They never view the same patch of Earth at the same time.]

The final requirement is that the metadata database be large, which impacts all transactions of the database. Compare MISR with 1 product per 150 minutes to MODIS with 1 product per 5 minutes and Landsat with 1 scene every 30 seconds. It would be expected that MODIS would have 30 times, and Landsat 300 times as much metadata as MISR.

Solution: meta-metadata

The solution for the CSP was meta-metadata in the form of models. The models provide information in a more compact form than can be provided through metadata. Once coincidences are found through the models, the metadata still must be searched because the models identify only a theoretical coincidence as opposed to a coincidence for which data were successfully captured and stored in the archive. The models enable coincidence search, as does the use of retail metadata, but without the same large volumes of data stored online. Coincidence search on products with wholesale metadata without meta-metadata requires detailed knowledge of the formats of the data to read the data in the archives. In contrast, the models are more abstract, so that data that is in a common format for all spacecraft must be used.

Coincidence search models

There are three models for coincidence search. The orbital model relates space and time by using equations for spacecraft dynamics to calculate the spacecraft position at times for which there is no measured value. The attitude model identifies the orientation of the spacecraft as a function of time. For Earth-observing spacecraft, this may simply be a constant value: the spacecraft always has one particular face oriented towards the Earth. The final model is the instrument model, which relates what an instrument views to the spacecraft position and attitude.

The orbital model requires information as to the state of the spacecraft at a given time, a means of computing estimates of the state at other times, and a way to control the growth of errors in the estimation of the state. Physics requires that the state include time and at least six components because there are at least six independent

Table 1. Approximate number of MODIS and Landsat products viewing an area of interest.

Time Duration	Size of Area of Interest (AOI)	MODIS Products		Landsat Products	
		Viewing AOI	Total	Viewing AOI	Total
30 days	Kansas sized	82	8,640	5.4	86,400
30 days	Continental U.S sized	212	8,640	1335	86,400
30 days	World	8,640	8,640	86,400	86,400
365 days	Continental U.S sized	2,579	105,120	16,242	1,051,200

variables. Typically, the orbital model starts with the spacecraft in a known state and integrates the forces (principally gravity and atmospheric drag for low-earth spacecraft) on the spacecraft to compute the estimate of the change in state of the spacecraft. Integration of the estimated changes in state produces the estimate of the state as a function of time. Alternatively, two times when the values of the state are known could be used to interpolate the estimate of the state between the times. Uncertainties in the model will cause errors in the estimation of the state, and if these errors are too large, the estimate is useless. If the demands for accuracy in the orbit state are not too severe, error growth can be controlled by limiting the timespan over which the orbital state is integrated before using a new known value for the orbital state. A second alternative is to use an analytic orbit algorithm. With an analytic method, the equations are handled (at least partially) analytically rather than through numerical integration. For example, for a given region of interest, and for a given instrument on a given spacecraft, it is possible to determine values of the longitude of the ascending node (a parameter of the orbit) such that the instrument will view the target area within the next orbit. Analytic propagation has sufficient accuracy to determine the longitude of the ascending node over a long timespan, which could be used to determine whether the instrument viewed the area of interest during a given orbit.

Fortunately, the orbit accuracy required for coincidence search is low, and the orbits of interest to EOSDIS are extremely predictable. Orbital errors tend to be the greatest along the path of the spacecraft (the along-track error). While along-track errors of 300 km are considered relatively large, for a spacecraft traveling at 7.7 km/s they correspond to an error in time of less than 40 seconds; hence, they are comparable to using the Landsat metadata and are a significant improvement over the MODIS metadata. The missions of interest to EOSDIS have circular orbits and, with the exception of the Tropical Rainfall Measuring Mission (TRMM), are at an altitude such that atmospheric drag is relatively insignificant.

The attitude model could be similar to the orbital model, but in practice is much simpler. It could be necessary to identify the attitude state at a given time and integrate the torques on the spacecraft to compute changes to the attitude state. Usually, however, the spacecraft autonomously controls the attitude state, so there is no need to know a measured value of the attitude. To a high accuracy, the commanded attitude state is the attitude state.

The instrument model relates what an instrument views to the orbital and attitude state. Usually, this means specifying how the instrument is mounted on the spacecraft and determining the instrument's field of view. Some instruments can vary their field of view relative to the spacecraft attitude (they are said to be "steerable").

Use of models in coincidence search data reduction

The models can be used to identify what timespans should include interesting data and are of more use if the timespans include the most interesting data that are identified. The timespans of interest can be when one instrument views a point or region or when multiple instruments view the same point or portion of a region. The timespans can be used to select only the products desired or additionally as input to subsetting routines to further reduce the selected products. If an assessment of the coincidence is provided to the user, the user can further reduce the data requested by selecting only the most interesting data.

The first way the models can be used to reduce the volume of data ordered by users is to allow users to order only the products they want. The models can be used to step through the time interval of interest and determine when the conditions for coincidence are met. The conditions for coincidence depend on the position (orbit), pointing (attitude), and field of view (instrument definition) of the instruments, which are obtained from the models using a small quantity of meta-metadata. Once coincidence times are identified, the metadata database(s) may be searched to determine whether the data are actually within the archive and how to obtain the data.

The second way the data volume is reduced is by supporting subsetting. Since the exact times when the instrument views the area of interest is determined as part of the coincidence identification, it can be used by subsetting routines. Lower-level spacecraft products tend to be organized by time and not necessarily in space. The EOS/AM-1 instrument Clouds and Earth's Radiant Energy System (CERES) produces products that are 24-hours long that do not include information to subset spatially. If the user specifies a spatial region for coincidence search and time range can be determined, that time range could be used to subset the data.

The third way to reduce data volume is by giving the user a way to select the best data. The user specifies search criteria that may provide more results than the user wants. If the results are returned with a quantitative assessment (see the glossary below) of each of the coincidence sets, the user may select only the best results. For example, the user can be provided with a measure of the area covered in the coincidence so that the user may select only those results where there is a large region of coincident data.

The fourth way to reduce data volume is by allowing the user to restrict selection of data based on orbital parameters not included in the metadata. Generally, Earth-viewing instruments are the most accurate in viewing the point on the Earth directly beneath them (the subsatellite point) and are the least accurate in viewing points on the horizon of the Earth (the limb). The user might wish to restrict the selection of data to only the data collected within some restricted radius of the subsatellite point.

The Coincidence Search Prototype

The CSP is a project for the Earth Science Data Information System (ESDIS) at GSFC that uses models to perform coincidence search. It is intended to support the EOSDIS V0 system and its successor, the ECS. To speed development of the system, Satellite Tool Kit (STK), a commercial off-the-shelf (COTS) product, was used for implementing the models. The information used by the CSP in place of metadata is 1 instrument definition file for each instrument, a database of orbit data with 1 record of data per spacecraft every 2 or 3 days, and the orbital, attitude, and instrument models.

CSP orbital model

The orbital model for the CSP uses the North American Defense Command (NORAD) Merged Simplified General Perturbations 4 (MSGP4) propagator, which takes NORAD two-line element sets (TLEs) as input. TLEs are available via the Web from a number of sources (ultimately all come from NORAD) for a variety of objects, including operational and no longer operational spacecraft for all nations. TLEs are available every 2 or 3 days, so that 10 to 15 TLEs (records) cover an entire month. Historical orbit data are available so that data collected during old spacecraft missions may be supported. Spacecraft position is obtained as a function of time by propagating the position of the spacecraft from the start time of one TLE until the start time of the next TLE.

CSP attitude model

The CSP attitude model assumes that the attitude of the supported spacecraft maintains a given orientation relative to the surface of the Earth. The model does not accept any input for changes to the attitude.

CSP instrument model

The CSP instrument model obtains input about the size, shape, and pointing of the field of view and constraints on operations. The size and shape of the field of view are defined as a circular, annular, elliptical, rectangular, or arbitrary simple closed polygonal shape. The size and shape are defined by angles, so the size of the field of view when projected onto the Earth varies with the altitude of the spacecraft. The pointing information is the direction of the center of the field of view of the instrument relative to the attitude as specified by the attitude model. Constraints on operations identify times when the instrument will not be in operation, usually by identifying limits on angles. The simplest example of such a constraint is only collecting data while viewing a region of the Earth in sunlight. If there are times when the instrument is known not to be in operation, the use of constraints can reduce the number of false coincidence sets identified for which queries must be made in the metadata database.

Operation of the CSP

The architecture of the CSP is diagrammed in Figure 2. It can be described as follows:

- In the background, the Query Manager polls an existing NASA file transfer protocol (FTP) site that provides orbit data in the form of TLEs. When new TLEs are found, they are stored in the TLE database.
- When the user wants to make a coincidence query, the user accesses the system through a Java-compliant browser. When the user submits a query, the Query Manager receives the query and provides the query with and necessary orbit data (TLEs) to the Space-Time Tool. The Space-Time Tool uses a COTS product, STK, to calculate coincidences. STK is run with two add-on modules: Coverage and Connect. Coverage supports the identification of space-time coincidences, including a quantitative assessment of the coincidence. The quantitative assessment is provided to the user, so one coincidence may be identified as better than another. Connect allows STK to interact with another program, rather than with a user, through the STK user interface. Logic within the Space-Time Tool extends the capabilities of STK to support the calculation of coincidences given a time tolerance.
- The Query Manager generates queries formatted as ODL based on results from the Space-Time Tool. For each coincidence pair found, two queries (one for each instrument) is generated. The queries are sent to an ODL-compliant metadata search interface. Currently, the only ODL-compliant interface is the V0 gateway, which is used to search archives at Langley Research Center (LaRC) and the Global Hydrology Research Center (GHRC) at the University of Alabama at Huntsville (UAH). The CSP will search ECS when ECS becomes operational.
- The Query Manager collects and reformats query responses from the ODL-compliant metadata search interface and provides the responses to the user with the qualitative assessment for each coincidence (from the Space-Time Tool) and information about how to order the data (from the ODL-compliant metadata search interface). Information is provided about what part of the data are of interest to the user, but currently there is no capability to subset the data.

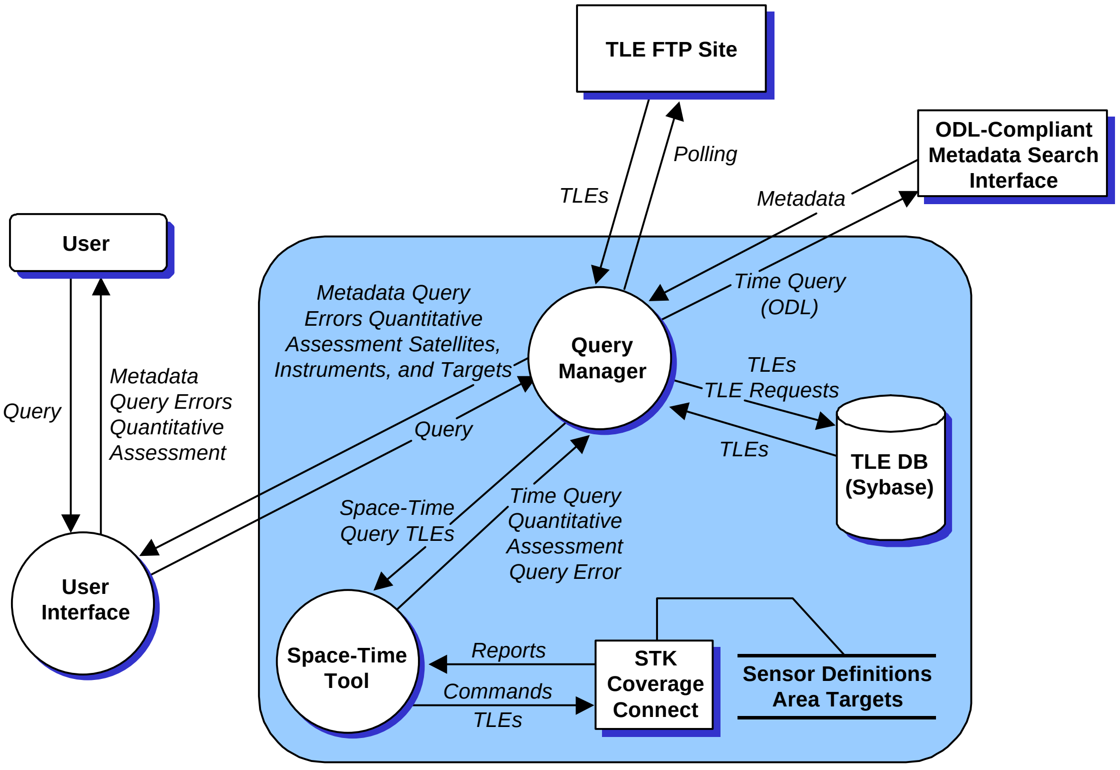

TLE FTP Site
Polling
TLEs
ODL-Compliant Metadata Search Interface
Metadata
User
Time Query (ODL)
Metadata Query Errors Quantitative Assessment Satellites, Instruments, and Targets
Query Manager
Query
Metadata Query Errors Quantitative Assessment
TLEs
TLE Requests
TLEs
TLE DB (Sybase)
Query
Space-Time Query TLEs
Time Query Quantitative Assessment Query Error
User Interface
Space-Time Tool
Reports
STK Coverage Connect
Commands
TLEs
Sensor Definitions Area Targets

Figure 2: Architecture of the CSP.

Summary

If the only tool you have is a hammer,
every problem looks like a nail.

Meta-metadata is an additional tool that can be used to enhance the capabilities of an archive while reducing the load on the archive. Meta-metadata is an abstraction of the metadata that supports more complicated queries than the metadata itself. Providing users with the capability to make more complicated queries allows the user to use the archive to separate the wheat from the chaff, so the user retrieves only the desired data. It was shown in this paper how ECS metadata was insufficient to provide spatial-temporal coincidence search, but that the problem could be solved using meta-metadata. Finally, the design of a prototype that implemented coincidence search using meta-metadata was described.

Glossary

Area Target: An area target is a definition of a geographic region by a geographic or phenomena name.

Coincidence Search: Coincidence search is the retrieval of metadata for groups of granules (two or more) that have common parameters. Spatial-temporal coincidence search is the retrieval of metadata for groups of granules that have common space and time extent.

Coincident Set: A reference granule and the granules coincident with it.

Data Center: A facility for storing, maintaining, and providing data sets for expected use in ongoing and/or future activities. Data centers provide selection and replication of data and needed documentation and, often, the generation of user-tailored data products.

Data Granule: A product stored in the archive for which there is a searchable metadata record in the metadata database.

Ephemerides: Time-tagged position data for instrument platforms such as spacecraft, aircraft, or ships.

Groundtrack: The projection of an orbit onto the surface of the Earth. The latitude and longitude for a point on the groundtrack is the same as the latitude for a particular point in the orbit.

Quantitative Assessment (of Coincidence): The space-time tool using STK will compute a quantitative assessment of coincidences to aid the user in selecting the best coincidences. Implementation options include length of time of overlap, maximum area of overlap, and integral of area over time of overlap. Higher quantitative assessment values would imply “better” coincidences.

Query Error: The query manager will check queries for nonsensical queries and return a query error to the user interface in response to a nonsensical query. Additionally, if the space-time tool determines that there is no coincidence given the user-specified constraints, the space-time tool will return a query error.

Sensor Definitions: Fields of view and pointing information about instruments.

Space-Time Query: A query that includes a spatial and a temporal constraint.

Time Query: A query that has a temporal constraint, but no spatial constraint.

A note on calculations

This subsection is to explain calculations made for this paper. For illustrative purposes through this paper, values were given for such things as the probability that an instrument would view an area of a certain size in a set period of time. The actual values for such things depend on the latitude of the area to be viewed, the shapes of the area to be viewed, and the field of view of the instrument. In the absence of that information, simple approximations were made to give order of magnitude results to provide insight into the problem. These approximations are admittedly crude and are not part of the algorithm used by the CSP. The chain of calculations is as follows:

1. The area covered by a product is calculated as the width of the swath time the length of the swath.
2. The total area viewed by the instrument is calculated. If the instrument views the entire Earth, the area is calculated as the 2 $(Re)^2$ where Re is the radius of the Earth and = 3.14159…. If the instrument views only between ±X latitude, the area at each pole that is not covered by the instrument is 2 (1 -cos(90-X))$(Re)^2$.
3. The probability that a point on the Earth is covered by a product is computed as the ratio of the results from calculations 1 and 2. This is true, on average, but is not true for a specified latitude. This would be accurate if the instrument viewed the surface of the Earth uniformly. In fact, the Equator is viewed least, and the latitude extremes are viewed the most.
4. The probability that a region on the Earth is viewed is computed as the area covered by the product (calculation 1) plus the area of the region entire quantity divided by total area viewed by the instrument (calculation 2). In the case that the area viewed is a point, this reduces to calculation 3. Effectively, the field of view of the instrument is expanded by the area of the region to be viewed.
5. The number of products generated in a timespan (for example, 30 days) is calculated as the timespan divided by the duration of the product.
6. The number of products that view a point (or a region) in a timespan is calculated as the number of products in that timespan (calculation 5) times the probability that the point (or region) is covered by one of the products.
7. The probability that during a given period of time a product from one instrument views some part of the same region at the same time that there is a product for another instrument that views an area is calculated as the duration of the first product plus the duration of the second product times calculation 6 for the second instrument divided by the duration of the period of time. This assumes that the orbits for the two instruments are independent of each other. This is a false assumption for a pair of instruments selected from Landsat 7 and EOS-AM-1. The orbits of the two spacecraft are such that the two spacecraft have the same groundtrack, but shifted in time by approximately 15 minutes. For these two spacecraft, the probability that instruments on both will view the same point on the ground simultaneously is 0.

Acknowledgements

Darrell Roos, Louis Reich, and Joan Dunham (CSC) participated in the concept definition of the CSP and in refining the architecture. A build of the CSP was successfully demonstrated following implementation by Sergey Nikhinson and Raj Kotaru (CSC). Louis Reich assisted in drafting the abstract for this paper. Joan Dunham and John Jensen (National Climatic Data Center) provided review comments.

References

[1] B.R. Barkstrom, “Digital Archive Issues from the Perspective of an Earth Science Data Producer,” *Digital Archive Directions Workshop,* June 1998

[2] “Coincidence Search Tiger Team Forum” at http://observer.gsfc.nasa.gov/cdwg/coincident/forum.html

[3] “Coincident Search Tiger Team Final Recommendations” at http://ecsinfo.hitc.com/cdwg/coincident/final.html

[4] “Response to ESDIS Paper from 2/13/96: Coincidence Searching/Coincidence Assessment” available at http://ecsinfo.hitc.com/cdwg/coincident/background.html